Supplementary information for:

Machine learning to predict anti-TNF drug responses of rheumatoid arthritis patients by integrating clinical and genetic markers

Yuanfang Guan[1*†,] Hongjiu Zhang[1†^], Daniel Quang[1], Stephen C.J. Parker[1], Dimitrios A. Pappas2[,3] Joel M. Kremer[3,4], Fan Zhu[5]

[1] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA 48109

[2] Columbia University College of Physicians and Surgeons, New York, USA

[3] Corrona LLC Waltham, MA, USA

[4] Albany Medical College and The Center for Rheumatology. Albany, USA

[5] Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing, China 400000

^ Current affiliation: Microsoft, Inc, Seattle, WA, USA

† YG, FZ, and HZ equally contributed to the work.

* To whom correspondence should be addressed: gyuanfan@umich.edu

Supplementary Figures

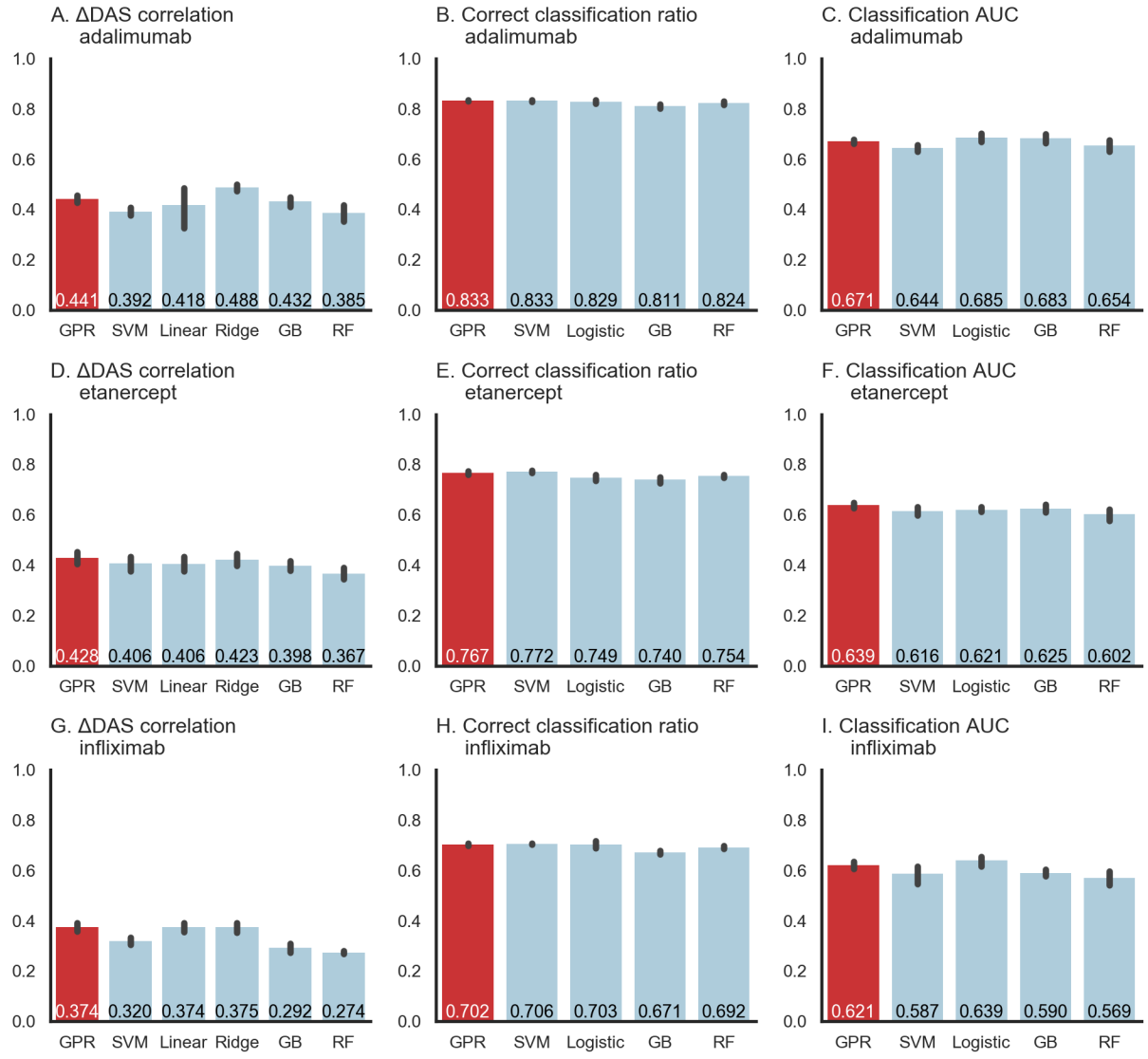

Figure S1. The cross-validation performance of common machine learning methods for individual anti-TNF drugs. The average scores are labeled above the X-axis. The final model is colored in red. (A, D, G) Pearson correlation coefficients between the observed ΔDAS and predictions from tested regression methods for adalimumab, etanercept, and infliximab. (B, E, H) Correct classification ratio of predictions from tested responder-vs-nonresponder classification methods for adalimumab, etanercept, and infliximab. (C, F, I) Areas under receiver operating characteristic curve (AUC) of tested responder-vs-nonresponder classification methods for adalimumab, etanercept, and infliximab. (GPR = Gaussian process regression, SVM = support vector machine, GB = gradient boosting regression/decision tree, RF = random forest)

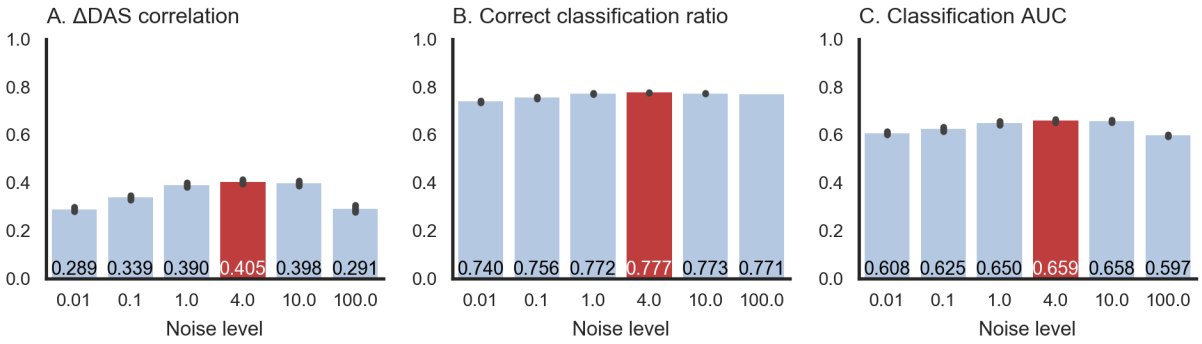

Figure S2. The cross-validation performance of different noise levels (α) specified in the GPR model. The average scores are labeled above the X-axis. The final model is colored in red. (A) Pearson correlation coefficients between the observed ΔDAS and predictions from tested noise levels. (B) Correct classification ratio of predictions from tested noise levels. (C) Areas under receiver operating characteristic curve (AUC) of tested noise levels.

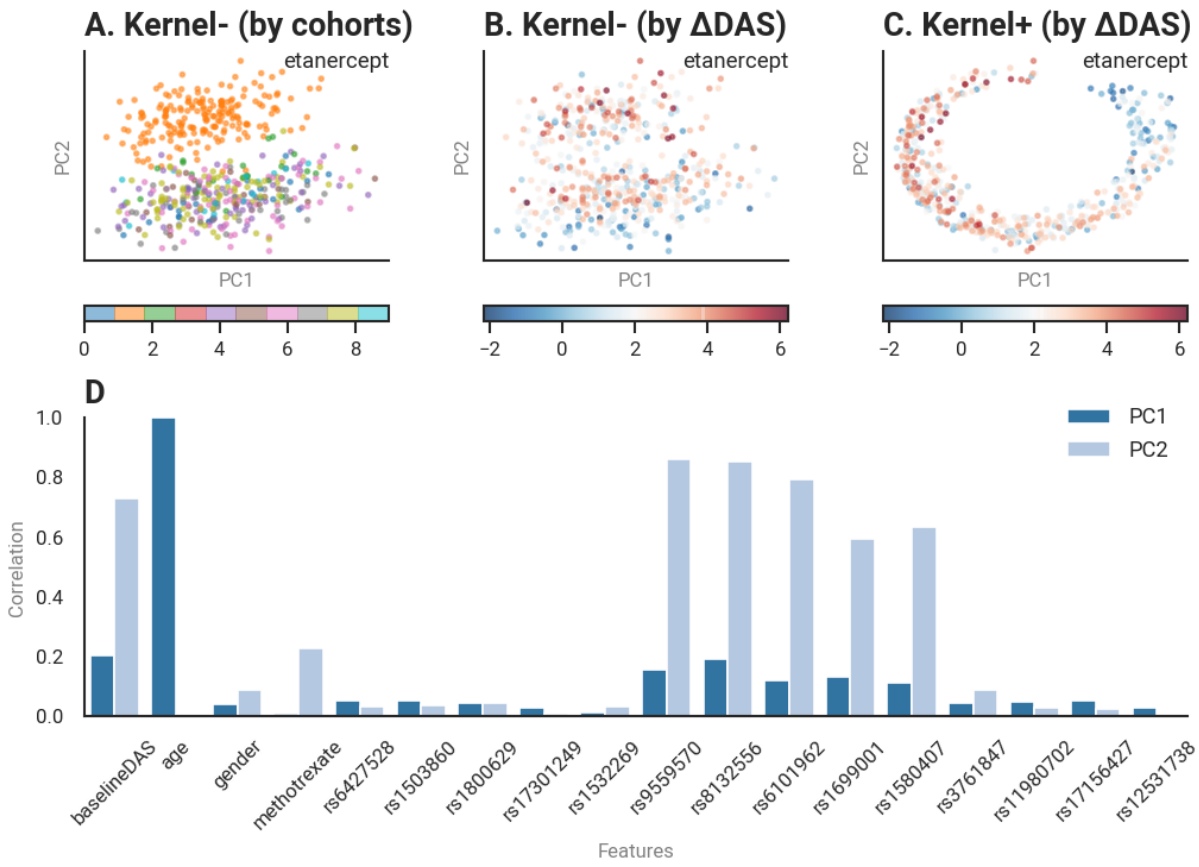

Figure S3. Feature space analysis of etanercept users in the training dataset. (A) Principal component analysis of the original feature space (without kernel transformation, colored in cohort labels) shows separation of several cohorts. (B) Principal component analysis of the original feature space (without kernel transformation, colored in ΔDAS) does not show obvious separation of responders and nonresponders. (C) Principal component analysis of the kernel matrix (colored in ΔDAS) shows a clear gradient from responders to nonresponders. (D) Feature contributions to first two principal component in Subfigure C.

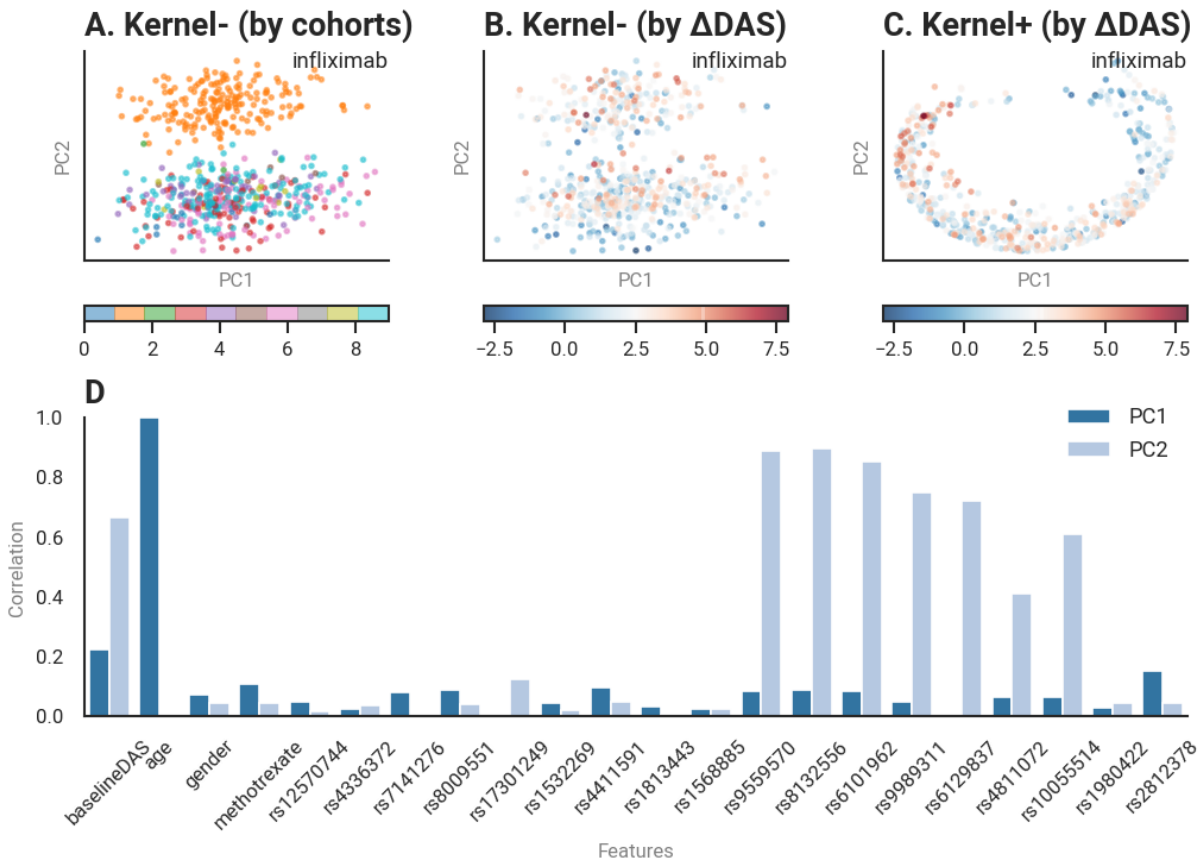

Figure S4. Feature space analysis of infliximab users in the training dataset. (A) Principal component analysis of the original feature space (without kernel transformation, colored in cohort labels) shows separation of several cohorts. (B) Principal component analysis of the original feature space (without kernel transformation, colored in ΔDAS) does not show obvious separation of responders and nonresponders. (C) Principal component analysis of the kernel matrix (colored in ΔDAS) shows a clear gradient from responders to nonresponders. (D) Feature contributions to first two principal component in Subfigure C.

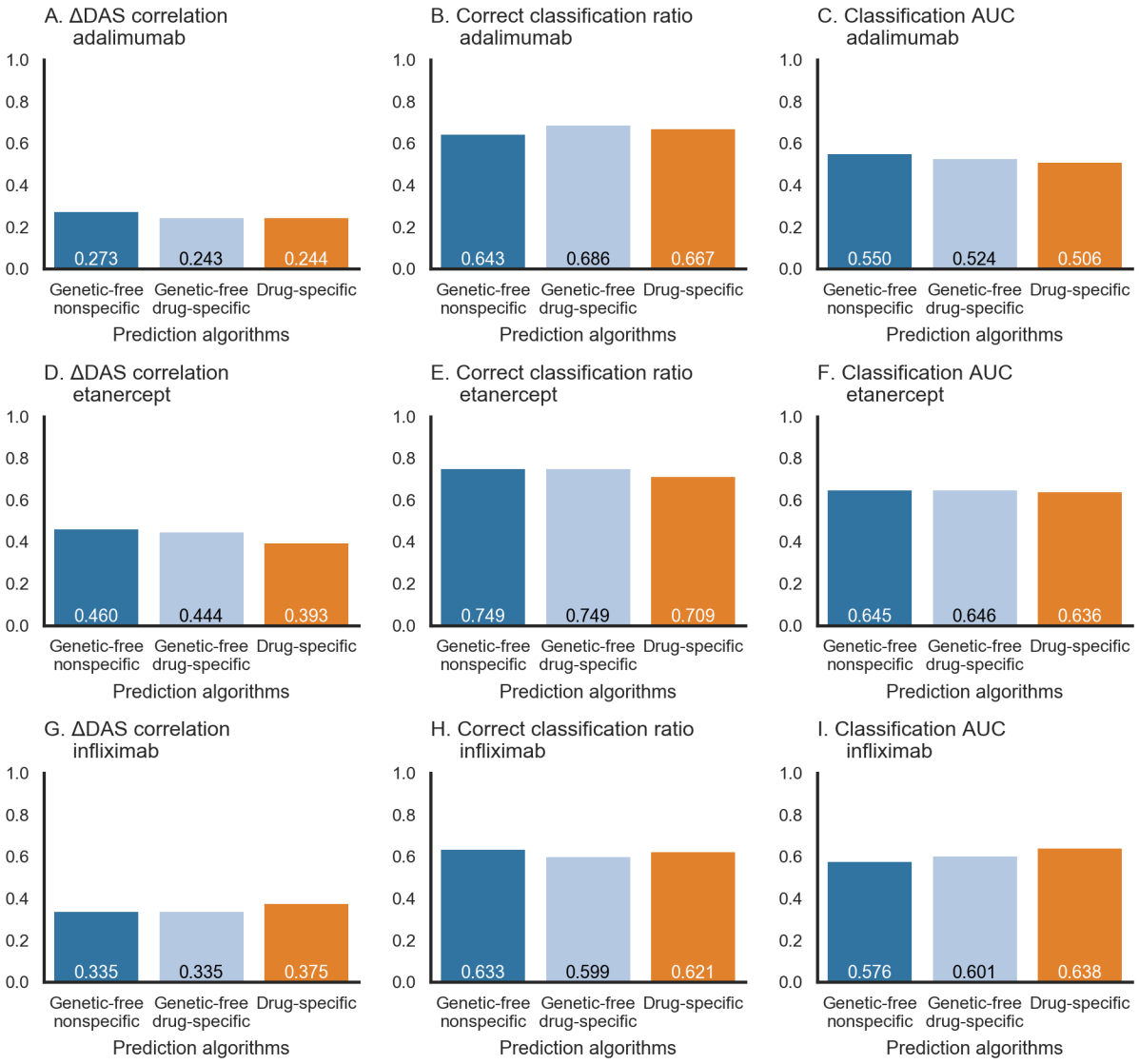

Figure S5. Evaluation of drug-specific models and non-specific models on the CORRONA dataset. The average scores are labeled above the X-axis. The final model is colored in red. (A, D, G) Pearson correlation coefficients between the observed ΔDAS and predictions from tested regression methods for adalimumab, etanercept, and infliximab. (B, E, H) Correct classification ratio of predictions from tested responder-vs-nonresponder classification methods for adalimumab, etanercept, and infliximab. (C, F, I) Areas under receiver operating characteristic curve (AUC) of tested responder-vs-nonresponder classification methods for adalimumab, etanercept, and infliximab.

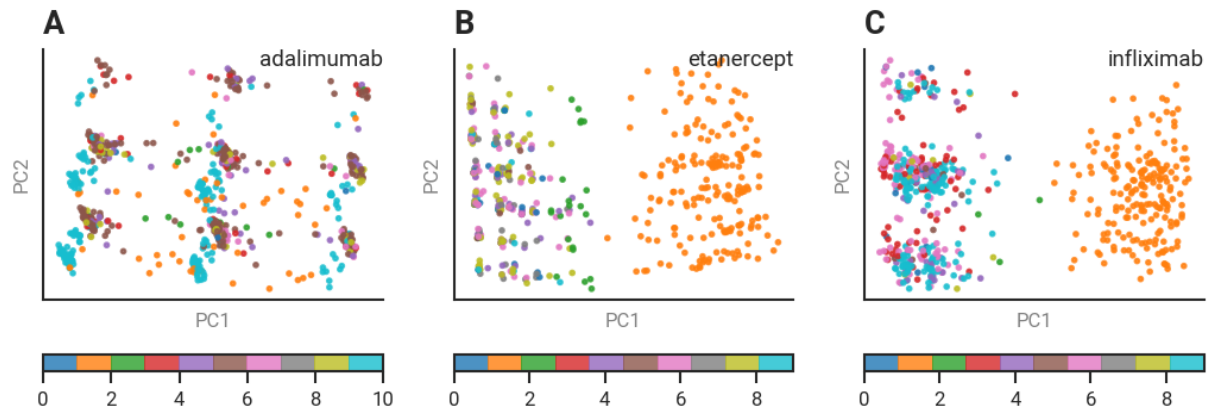

Figure S6. The principal component analysis on genetic features in the challenge training dataset. The dataset was divided based on the drugs. The colors corresponds to the cohort indices. The clustering of samples from the same cohort demonstrates the association of the genetic features to the cohort information.

Supplementary Tables

Table S1. Bootstrap test evaluation of various prediction algorithms.

Classification methods (The numbers of bootstrap rounds that GPR outperformed other algorithms in listed metrics)		
	Accuracy (%)	Area under ROC
Gradient boosting	95	46
Logistic regression	82	78
Ridge regression	91	82
Random forest	100	93
Support vector machine	64	97
Regression methods (The numbers of bootstrap rounds that GPR outperformed other algorithms in listed metrics)		
	Pearson correlation coefficient	
Gradient boosting	75	
Linear regression	53	
Ridge regression	46	
Random forest	100	
Support vector regression	100	

Table S2. Genetic markers included in the GPR model for predicting anti-TNF drug responses. The gene is collected from dbSNP GeneView report. The correlation value is the Pearson correlation coefficient between the SNP dosage and the patients' ΔDAS. Fisher F-test was performed on the training dataset against the response classification. Fast-LMM p-value is calculated by Microsoft Fast-LMM in all mode, with all clinical data and treatment present and deltaDAS as phenotype. Literature search were done through PubMed using keywords of "rheumatoid arthritis", "anti-TNF", "response" and "marker".

SNP	Gene	Correlation	Fisher F-test	Fast-LMM p-value	Literature	Function
Adalimumab						
rs10265155	MAGI2	0.0250	0.0134	0.392	PMID 23555300	Guanylate kinase
rs1990099	MAGI2	0.0251	0.0135	0.403	PMID 23555300	Guanylate kinase
rs10833455	NELL1	0.00817	0.0877	0.696	PMID 23555300	Protein kinase C binding
rs10833456	NELL1	0.00813	0.0854	0.696	PMID 23555300	Protein kinase C binding
rs7932820	NELL1	-0.00787	0.00831	0.687	PMID 23555300	Protein kinase C binding
rs17301249	EYA4	0.0230	0.179	0.465	PMID 21061259	Transcription co-activator and phosphatase
rs1532269	PDZD2	0.0328	0.543	0.006	PMID 21061259	Unclear
rs4411591	LINC01387	0.00594	1.00	0.001	PMID 23233654	Unclear
rs1813443	CNTN5	-0.0394	0.549	0.005	PMID 23233654	Immunoglobin
rs1568885	LOC107986770	0.0145	0.747	0.076	PMID 23233654	Unclear
rs940928	EDAR	-0.0199	0.0187	0.926	PMID 23555300	Ectodysplasin receptor
rs12226573		0.0166	0.0149	0.036	US20170145501A1	
rs7933314		-0.0362	0.180	0.132	US20170145501A1	

rs8132556	FAM3B	-0.176	0.286	0.866		Cytokine-like
rs9559570		0.221	0.186	0.806		
rs621213	LOC107985260	-0.156	0.170	0.433		Unclear
rs620336	LOC107985260	0.156	1.00	0.433		Unclear
rs1980422		-0.0269	0.922	0.972	PMID 23007924	
rs2812378	CCL21	-0.002669	1.00	0.104	PMID 20461788	Cytokine
Etanercept						
rs6427528	CD84	0.0699	0.151	0.045	PMID 23555300	Self-ligand receptor of the signaling lymphocytic activation molecule
rs1503860	CD84	-0.0713	1.00	0.037	PMID 23555300	Self-ligand receptor of the signaling lymphocytic activation molecule
rs1800629	TNF	0.0392	0.608	0.403	PMID 19365401	tumor necrosis factor (direct target)
rs17301249	EYA4	-0.0222	0.528	0.465	PMID 21061259	Transcription co-activator and phosphatase
rs1532269	PDZD2	-0.0369	0.598	0.006	PMID 21061259	Unclear
rs9559570		0.455	0.0965	0.806		
rs8132556	FAM3B	0.438	0.0159	0.866		Cytokine-like
rs6101962		-0.354	0.0922	0.434		
rs1699001	BTBD9	-0.334	0.0362	0.866		Unclear
rs1580407		-0.312	0.860	0.653		

rs3761847	TRAF1	-0.00333	0.501	0.896	PMID 17804836	TNF receptor associated factor (direct target)
rs11980702		-0.0104	1.00	0.770	US20170145501A1	
rs17156427		0.00832	1.00	0.430	US20170145501A1	
rs12531738		-0.0178	0.459	0.412	US20170145501A1	
Infliximab						
rs12570744		0.0264	0.00313	0.078	PMID 23555300	
rs4336372		0.0164	0.254	0.913	PMID 23555300	
rs7141276		0.0274	0.0263	0.189	PMID 23555300	
rs8009551		-0.0452	0.292	0.180	PMID 23555300	
rs17301249	EYA4	0.0830	0.296	0.465	PMID 21061259	Transcription co-activator and phosphatase
rs1532269	PDZD2	0.0189	0.647	0.006	PMID 21061259	Unclear
rs4411591	LINC01387	0.0118	1.00	0.001	PMID 23233654	Unclear
rs1813443	CNTN5	-0.0442	0.122	0.005	PMID 23233654	Immunoglobin
rs1568885	LOC107986770	0.00782	0.120	0.076	PMID 23233654	Unclear
rs9559570		0.0407	0.0155	0.806		
rs8132556	FAM3B	-0.400	0.00506	0.866		Cytokine-like
rs6101962		-0.359	0.727	0.434		
rs9989311	LOC646214	-0.352	1.00	1.0		Unclear
rs6129837	CHD6	0.324	0.233	0.708		Transcription repressor
rs4811072		-0.273	0.0304	0.049		

rs10055514		0.288	0.0157	0.964		
rs1980422		0.00698	0.208	0.972	PMID 23007924	
rs2812378	CCL21	0.0862	0.616	0.104	PMID 20461788	Cytokine

Table S3. Pearson correlation coefficients between baseline DAS and ΔDAS.

	Total	Adalimumab	Etanercept	Infliximab
Training cohort	0.370	0.418	0.407	0.314
CORRONA cohort	0.351	0.206	0.477	0.359